Eyewitness descriptions without memory: The (f)utility of describing faces

Robin S. S. Kramer and Georgina Gous

School of Psychology, University of Lincoln, UK

Corresponding Author:

Robin Kramer, School of Psychology, University of Lincoln, Lincoln LN6 7TS, UK.

E-mail: remarknibor@gmail.com

Telephone: +44 (0)1522 835806

Running head: Eyewitness descriptions without memory

Conflicts of interest

There are no conflicts of interest to be disclosed.

Acknowledgements

The authors thank Sarah Pinder, Caitlin Barber, and Joanna Thompson, as well as our Research Skills IV students, for collecting the data.

Abstract

Eyewitness descriptions provide critical information for the police and other agencies to use during investigations. While researchers have typically considered the impact of memory, little consideration has been given to the utility of facial descriptions themselves, without the additional memory demands. In Experiment 1, participants described face images to their partners, who were then required to select these faces from photographic lineups. Performance was error-prone when the same image appeared in the lineup (73% correct), and decreased further when a different image of the same face was presented (22% correct). We found some evidence to suggest this was due, in part, to difficulties with recognising that two different images depicted the same person. In Experiment 2, we demonstrated that descriptions of the same face given by different people showed only moderate agreement. Taken together, these results highlight the problematic nature of facial descriptions, even without memory, and their limited utility.

Keywords

Eyewitness; facial descriptions; memory; face matching; agreement

1 Introduction

Eyewitness descriptions often play a substantial role in criminal investigations (Brown, Lloyd-Jones, & Robinson, 2008). While initially useful in locating suspects immediately after an incident has taken place, by issuing a "be on the lookout" (BOLO) or through other means, these descriptions also provide critical information for the police to use during criminal investigations. For example, descriptions provided by eyewitnesses may drive the identification of potential suspects from mug books or the construction of suspect sketches/composites circulated to the public (Davies, 1981; Meissner, Sporer, & Schooler, 2007), the selection of fillers for use in live or video lineup identification parades (Kebbell, 2000), and the subsequent assessment of fairness for those lineups (Meissner et al., 2007). Witness descriptions are also frequently introduced at trial as a means of demonstrating the congruence between the suspect's testimony and that of the witness (Meissner et al., 2007). While there are several factors that can influence the accuracy and utility of these descriptions, research to date has yet to consider one core aspect – our ability to produce facial descriptions. Simply, are people able to describe faces in a way that is sufficient for identification?

Since eyewitness descriptions are generated following, and not during, an incident, it is understandable that researchers have predominantly focussed on the role of memory. Unsurprisingly, evidence has shown that delay between exposure to a person and subsequently describing that person has significant detrimental effects on both accuracy and completeness of descriptions (e.g., Ellis, Shepherd, & Davies, 1980; Meissner, 2002; van Koppen & Lochun, 1997). However, it is worth noting that the passage of time alone may not be the underlying cause of these impairments, and that both the strength of the initial memory trace and the nature of any interference during the delay interval produce significant detriments. For example, low levels of illumination (DiNardo & Rainey, 1991; Yarmey, 1986), high levels of eyewitness stress and anxiety (Deffenbacher, 1994; Deffenbacher, Bornstein, Penrod, & McGorty, 2004), the presence of a

weapon (Pickel, 1999; Steblay, 1992), a shorter duration of exposure to the perpetrator (Yarmey, Jacob, & Porter, 2002), a greater distance between the witness and perpetrator (van Koppen & Lochun, 1997), and being under the influence of alcohol and drugs (Sporer, 1992; Yuille & Tollestrup, 1990; 1992; Yuille, Tollestrop, Marxsen, Porter & Herve, 1998) have all been found to impair memory. Researchers have also been able to influence the accuracy of eyewitness memory through exposure to post-event information (e.g., receiving information provided by another witness – Shaw, Garven, & Wood, 1997).

In addition, witness characteristics can affect the accuracy of person descriptions, with women demonstrating better memory for crimes than men (e.g., Lindholm & Christianson, 1998), and younger (versus older – Yarmey, 1993) adults and older (versus younger – Davies, Tarrant & Flin, 1989) children showing more accurate eyewitness recall. Differences between ethnicities have also been noted, with research suggesting that individuals attend to features deemed relevant to faces of their own ethnicity, applying this inappropriately when examining faces of other ethnicities. For example, Ellis, Deregowski, and Shepherd (1975) found that Black and White participants recalled different features when describing faces, with the latter group often reporting more redundant information when describing Black faces (e.g., "he has black skin, black kinky hair and brown eyes" – p.123).

In one influential paper, Megreya and Burton (2008) were able to demonstrate that factors other than memory also contributed to poor eyewitness performance. In their first experiment, participants viewed a target for 30 s (either live or as a static image), followed by a 5 s gap, and were then presented with a 10-face image array. Crucially, in trials where the target was present in the array, the array image differed from the original photograph that was viewed to avoid simple picture matching. Under these idealised conditions (minimal memory requirement, no stress, high quality and front-on images, etc.), performance was poor – around 60% accuracy for target-present trials and 80% accuracy for target-absent trials. Following on from this, their second experiment removed memory altogether from the task, with the target person/image presented alongside the 10-

face image array. Even when participants only had to match the target to an array image, accuracy levels were around 70% (target-present) and 65% (target-absent) correct. This result highlights the difficulties that people have with simply comparing an unfamiliar person's face across instances, providing a conceptual baseline for eyewitness identification, where inherent memory demands represent an additional obstacle to accuracy.

Eyewitness identification is clearly limited in its accuracy, even when memory requirements are absent. As discussed earlier, eyewitness descriptions have also been shown to suffer because of the need to remember and then later describe a face. Whether through the delay itself or the experiences during the intervening period, the memory of the face is typically subject to degradation and noise. However, there is another important component that has been little studied to date – the process of describing the face. Being able to describe the remembered face is not only dependent on the accuracy of the representation in memory but also on the ability to convey the appearance of the face to others.

Through analysing archival data from actual police records, van Koppen and Lochun (1997) found that eyewitness descriptions tended to provide little information regarding the perpetrators, and typically included more general features (e.g., sex, race, build) rather than specific facial characteristics (see also Kuehn, 1974; Yuille & Cutshall, 1986). Of course, this may be due to the limitations in which kinds of information were available to the witness during the crime. If the offender had his or her back to the witness, for example, then this would prevent any facial description.

In both laboratory studies (Ellis et al., 1980; Laughery, Duval, & Wogalter, 1986) and archival data (Sporer, 1992), researchers have shown that verbal descriptions mentioned upper face features more frequently than features of the lower face, and predominantly involved exterior facial descriptors (e.g., hair details; Lindsay, Martin, & Webber, 1994; Pozzulo & Warren, 2003). For instance, in one study, participants focussed most often on iris colour, as well as hair colour and

texture, when required to provide a description of the target face to an experimenter (Ellis et al., 1975).

Evidence has shown that accuracy levels when recalling facial features in eyewitness descriptions may be relatively poor (van Koppen & Lochun, 1997). However, such findings are a combination of an (in)accurate memory of the face *and* the (in)ability to describe the face to a second person. Regarding the latter, "a common difficulty with person descriptions involves the limited vocabulary that individuals have for describing the human face" (Meissner et al., 2007, p. 16).

Recently, an experiment by Wilson and colleagues (2018) specifically considered the task of selecting a face from an eight-image array based solely on a written description provided by a previous participant (generated immediately after watching a video of a mock bank robbery). Using only a target-present lineup, the researchers found that participants correctly identified the perpetrator on only 14% of trials using this description (where chance guessing would produce 13% correct). Although the utility of the description would necessarily have depended on how clearly the perpetrator was captured in the video, the results demonstrate how incapable descriptions may be of conveying appearance to others.

Underlying the utility of facial descriptions in providing appearance information is the assumption that people agree on which descriptors are applicable for a given face. If a suspect's nose is considered to be "long" by the witness but not by those who subsequently read the description, it is unclear how the description can be of value in locating or identifying the perpetrator. Studies featuring the mock witness paradigm (descriptions of the perpetrator are generated, collated, and then used to determine the fairness of lineups – Doob & Kirshenbaum, 1973) provide some suggestion of difficulties with this type of agreement. Typically, a sample of participants unrelated to the main study view the perpetrator (for example, in a video or image) and give their descriptions, with a modal description comprising characteristics that are mentioned by at least 25% of the sample (Beresford & Blades, 2006; Dekle, Beal, Elliott, & Huneycutt, 1996).

Although not especially strict, this criterion appears to result in only general descriptions that fail to incorporate facial appearance. For example, "White male in his 20s, medium build, medium height with short brown hair" (Beresford & Blades, 2006, p. 1106), "the perpetrator is a white female, medium to average height (5'3"–5'6"), medium to average weight (120–130 lbs), has medium- to shoulder-length dark brown hair with a slight wave, and is 25-30 years old" (Dekle et al., 1996, p. 4), or "white man, early 20s, dark short hair, medium build" (Humphries, Holliday, & Flowe, 2012, p. 151). The apparent lack of overlap in the use of facial descriptors by at least a quarter of participants in these studies is suggestive of low agreement across the descriptions that were produced.

In the current set of experiments, we investigated the task of describing faces to others in the absence of any memory demands. Experiment 1A considered whether the freely-generated descriptions of faces produced by one person can be used by a second person to select the target from a lineup. In Experiment 1B, we compared these results to those produced in a simple face matching task, where the step of describing a face to someone else is removed, in order to determine whether faces that were difficult to convey through descriptions were also those that were difficult to match. Finally, Experiment 2 considered descriptions produced through the completion of pre-defined items and scales used in earlier work, investigating an important requirement of face descriptions: the nature of agreement across describers.

2 Experiment 1A – Face descriptions

In this experiment, we investigated whether people were able to accurately describe faces to others when memory was not required. As such, we focussed solely on the descriptive component of eyewitness accounts. We made use of the 'communication accuracy paradigm' previously investigated by Fallshore and Schooler (1995). Crucially, however, where those researchers obtained participants' descriptions of faces generated from memory (after a 5 min filler task) and

gave these to a new set of participants for use in target identification, our participants' descriptions were produced while viewing the face under consideration, therefore removing the memory component.

2.1 Method

2.1.1 Participants

One hundred and thirty students (87 women; age M = 20.5 years, SD = 3.8 years; 90.8% self-reported ethnicity as White) at the university took part in the experiment. All participants provided written, informed consent before the experiment, and also received both verbal and written debriefings upon completion. The university's ethics committee approved all experiments presented here, which were carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

2.1.2 Stimuli

For 24 identities (half women), two different facial photographs were downloaded from Google Images. These identities comprised White celebrities in Australia and Europe, chosen so that (a) two images could be found for each person; and (b) the people would be unfamiliar to our predominantly British participants. Care was taken to make sure that the two images depicted different situations (i.e., they were not taken minutes apart and so did not share clothing, backgrounds, etc.) but were otherwise unconstrained, varying in lighting, pose, expression, and so on.

For each identity, a single image for each of ten different people ('foils') was downloaded using descriptive search terms that matched the identity's general appearance (e.g., 'blonde woman'

and ‘grey haired man’). Foils were also chosen to be unfamiliar to our participants, and no foil was used for more than one identity. In total, this resulted in a final stimulus set of 288 images: 2 identity and 10 foil images for each of 24 identities.

In order to make sure that the face in each photograph was clearly visible, images met the following criteria: (a) no part of the face should be obscured (for example by clothing, glasses, or a hand); (b) pose should be very broadly full-face; and (c) pose should be standing or sitting, but not lying down, in order to limit the angle of the head to relatively upright. All images were high quality, colour, naturalistic photographs, and were cropped loosely around the person’s head and then resized to 3.3 x 5 cm.

2.1.3 Procedure

Participants were tested in pairs. The roles of ‘describer’ and ‘listener’ were randomly assigned at the start of the experiment, and these roles were reversed after half of the trials (i.e., 12) had been completed. Materials were printed in colour and given to participants as a paper booklet. On each trial, the describer was presented with a single image of the target identity and was instructed to describe the person to the listener. Mirroring the collection of real-world facial descriptions, no constraints were placed on this process, and so descriptions could include both featured-based (e.g., “big nose”) and more holistic statements (e.g., “looks attractive”). The listener was provided with a ten-image array (with images numbered from 1 to 10) and was asked to select the identity if they thought the person was present but could also respond “absent” if they did not think the person appeared in the array. The listener was not permitted to ask questions of the describer, nor were they allowed to prompt the describer for additional information etc., and so the describer simply continued to describe/elaborate until the listener gave their response orally. (No time constraints were imposed.) Participants were unable to see each other’s booklets during the experiment.

For each trial, the experimenter wrote down the listener's response (the number of the image selected or "absent"), as well as noting which (if any) particular features the describer mentioned from the following: eyes, nose, mouth, hair, ears, face shape, age, expression. If additional features were described frequently, the experimenter also made a note of these. No time constraints were imposed on the task.

The 24 identities/trials were allocated to one of three conditions (see Figure 1). In the 'same' condition, the image of the target identity being described also appeared in the ten-image array given to the listener (along with nine of the foil images). As such, both participants in the pair saw the same image. It is worth noting that the background of the image was therefore available for use by the describer, although the particular condition for any given trial was unknown to the pair. In the 'different' condition, one image of the target identity was described while the other image of that identity appeared in the listener's array (along with nine of the foil images). Finally, in the 'absent' condition, the identity being described did not appear in the array, with all ten images depicting foils. Where the target identity was present in the ten-image array, the image's position in the array was initially randomly selected for that trial and then was held constant across participant pairs.

The allocation of the trials to conditions, and the order of the 24 trials, were initially randomised with the proviso that the three conditions were equally represented in the first and second halves of the experiment, meaning that each of the participants in the pair described four trials for each condition. This trial allocation and order were then held constant across pairs.

2.2 Results

For each participant, we calculated the proportion of correct responses (out of four trials) for the three conditions separately. These proportions were analysed using a one-way (Condition: same, different, absent) within-subjects analysis of variance (ANOVA). We found a significant main

effect, $F(2, 258) = 124.26$, $p < .001$, $\eta_{2p} = 0.49$, with pairwise comparisons (Bonferroni corrected) indicating that all three conditions significantly differed from each other (all *p*s < .001). These differences are illustrated in Figure 2, with additional details presented in Table 1.

Given that the performance of each participant was inherently dependent on the other person in their pair, we carried out the same analysis at the level of pair. For each pair, we calculated the proportion of correct responses (out of eight trials) for the three conditions separately. We again found a significant main effect, $F(2, 128) = 90.15$, $p < .001$, $\eta_{2p} = 0.59$, with pairwise comparisons (Bonferroni corrected) indicating that all three conditions significantly differed from each other (all *p*s < .001). Indeed, given that each pair comprised two participants in the previous analysis, the condition means remained unchanged.

With performance in the 'different' condition being so low (see Figure 2), we also examined the errors people made for those trials. For each participant, on 'different' trials where an incorrect response was given, we noted whether the participant selected a foil or responded "absent". This resulted in a proportion for each of the two response types for each participant. Given that these values were the complement of each other (necessarily summing to 1), we compared one of these error types to a value of 0.5 (since both comparisons would produce the same statistical result). We found a significant difference, $t(129) = 2.26$, $p = .026$, Cohen's $d = 0.20$, demonstrating that the proportion of "absent" responses ($M = 0.56$) was larger than the proportion of responses in which a foil was selected ($M = 0.44$).

Next, we analysed how often participants mentioned particular facial features and whether this differed across conditions and/or was associated with answering correctly versus incorrectly. Table 2 summarises the frequencies with which each of the features was mentioned. At the level of pairs, we calculated the proportion of trials in which each of the facial features (eyes, nose, mouth, hair, ears, face shape, age, expression) was mentioned, separately for the three conditions and for correct versus incorrect responses. For each facial feature, these proportions were then analysed

using a 3 (Condition: same, different, absent) x 2 (Response: correct, incorrect) within-subjects ANOVA.

For the eyes, neither the main effects of Condition or Response nor the interaction between the two were statistically significant (all *p*s > .457, all η_{2p} < 0.02). This pattern of results was also found for the nose (all *p*s > .496, all η_{2p} < 0.02), the ears (all *p*s > .205, all η_{2p} < 0.04), the face's shape (all *p*s > .316, all η_{2p} < 0.03), age (all *p*s > .208, all η_{2p} < 0.05), and expression (all *p*s > .382, all η_{2p} < 0.02). For the mouth, we found a significant main effect of condition, $F(2, 70) = 3.70$, $p = .030$, $\eta_{2p} = 0.10$, although none of the pairwise comparisons remained significant after Bonferroni correction. Neither the main effect of Response nor the interaction were statistically significant (both *p*s > .303, both η_{2p} < 0.03) for the mouth. Finally, for the hair, we again found a significant main effect of condition, $F(2, 70) = 8.58$, $p < .001$, $\eta_{2p} = 0.20$, with this feature being mentioned more often during 'different' trials than 'same' ($p < .001$) and 'absent' trials ($p = .042$). These latter conditions did not differ from each other ($p = .361$). Neither the main effect of Response nor the interaction were statistically significant (both *p*s > .055, both η_{2p} < 0.08) for the hair.

We also carried out this same analysis for the total number of these features mentioned (out of eight) for each type of condition and response. We found a significant main effect of Response, $F(1, 35) = 4.50$, $p = .041$, $\eta_{2p} = 0.11$, with more features mentioned for incorrect ($M = 4.04$) in comparison with correct responses ($M = 3.80$). There was no main effect of Condition, $F(2, 70) = 2.50$, $p = .089$, $\eta_{2p} = 0.07$, and no interaction, $F(2, 70) = 0.66$, $p = .521$, $\eta_{2p} = 0.02$.

2.3 Discussion

The results demonstrated how difficult our participants found this task. Describing a face from a photograph to someone who has the same photograph in their array produced reasonable levels of accuracy (though still far from perfect). However, simply switching this image for a different

photograph of the same person resulted in a striking decrease in performance. Typically, this change of image led to the listener incorrectly responding that the target was absent from the array.

Regarding the descriptions used, we found no differences in the facial features mentioned by participants when comparing correct and incorrect trials. In other words, performance was not dependent on describing particular features of the face. However, perhaps surprisingly, we found that more features were mentioned on incorrect rather than correct trials. This may be due to the nature of the task. Given that describers continued talking until listeners made their decisions, it may be that this process took longer on trials in which responses were ultimately incorrect. Initial descriptions failed to lead the listener to the correct answer, resulting in the mention of additional features (but not the correct response). It is also worth noting that the features analysed here were limited to the list of eight that we coded for, which included age, hair, and expression. All three features, and the latter two in particular, often changed across images and so were not diagnostic of identity. As such, it may be that describing more features resulted in included those that were misleading to listeners, potentially producing incorrect identifications. We also note that participants discussed features not included in our list (e.g., eyebrows and teeth), and so these additional features may have improved response accuracy but were not measured here.

3 Experiment 1B – Face matching

Next, we aimed to determine whether trial accuracy in the first experiment was associated with face matching accuracy using those same images. If trials were more difficult when describing the face to a second person, could this be due, at least in part, to difficulties in simply matching the images when no description was required?

3.1 Method

3.1.1 Participants

One hundred and thirty volunteers living in the UK (60 women; age *M* = 29.4 years, *SD* = 10.7 years; 85.4% self-reported ethnicity as White) took part in the experiment. Recruitment took place through approaching people on campus or via Amazon Mechanical Turk. All participants provided written, informed consent in person or online before the experiment, and also received a written or online debriefing upon completion.

3.1.2 Stimuli

The same stimuli were used here as in Experiment 1A.

3.1.3 Procedure

Participants were tested individually in person or online. For those who participated in person, materials were printed in colour and given to participants as a paper booklet. For online participants, the experiment was carried out using the Qualtrics survey platform. On each trial, the participant was presented with a single image of the target identity, along with a ten-image array, and was asked to select the identity if they thought the person was present but could also respond "absent" if they did not think the person appeared in the array. No time constraints were imposed on the task.

Of the 24 trials in Experiment 1A, we removed the eight 'same' condition trials as these would be trivial in a face matching task. As such, participants were only presented with the remaining 16 trials (i.e., the 'different' and 'absent' conditions). Responses were either written down by the experimenter (in person) or provided using the mouse (online).

3.2 Results

First, we considered individual-level performance, with a summary of these data provided in Table 3. Performance levels were comparable with previous studies investigating matching using a 10-face image array procedure (e.g., Bruce et al., 1999).

Next, we considered trial-level performance. For each of the 16 trials, we calculated the proportion of correct responses across all participants. Using the data collected in Experiment 1A, we also calculated the proportion of correct responses across pairs of participants. In order to determine whether more difficult trials in terms of describing were also more difficult for matching, we then correlated these two sets of proportions. We found a moderate, although not statistically significant, association between the two measures, $r(14) = 0.36$, $p = .174$.

Although these data contained only eight trials from each condition, we also carried out the analyses separately for 'different', $r(6) = 0.64$, $p = .086$, and 'absent' trials, $r(6) = 0.22$, $p = .609$.

3.3 Discussion

Although cautious to draw any conclusions based on so few trials for each condition, the results provide some suggestion that in trials where it was difficult to describe the face to a listener in order to foster selection from the array, participants also found it difficult to match the target image to the second image of the identity in the array. Again cautiously, it appears that for trials where the target was absent from the array, there was less of a relationship between determining this in a describing/listening task and during a matching task.

4 Experiment 2 – Agreement in face descriptions

Finally, we investigated a key question when determining the utility of face descriptions: do describers agree with each other? If every person describes the same face in a different way then this places an upper boundary on the description's utility. Put simply, if Person A thinks the target's nose is long but Person B thinks it is short, there is no clear way that the description can be valuable in locating or identifying the target.

4.1 Method

4.1.1 Participants

Forty volunteers (27 women; age M = 34.3 years, SD = 18.7 years; all self-reported ethnicity as White), recruited through word of mouth, took part in the experiment. All participants provided written, informed consent before the experiment, and also received both verbal and written debriefings upon completion.

4.1.2 Stimuli

Four identities (two women) were selected at random from those presented in Experiment 1A. For each of these, the image previously used by the 'describer' was chosen for the current experiment.

4.1.3 Procedure

Participants were tested individually. Images were printed in colour and given to participants for consideration. Each participant was presented with a single face image and asked to complete the Face Rating Scales (FRS) questionnaire (Sporer, 2007), which was a modified version of the Aberdeen University Face Rating Schedule (Ellis, 1986). Assignment of participants to faces was

based on when they took part, cycling through the four faces, resulting in ten participants rating each face.

The FRS involved participants rating 42 facial characteristics on 5-point anchored scales (e.g., quantifying the eyebrows from '1 - narrow' to '5 - broad'), as well as providing 11 yes/no responses (e.g., the presence/absence of glasses). (For more details, see Sporer, 2007.) Upon completion of the questionnaire, participants provided demographic information.

4.2 Results

For the 42 items utilising rating scales, we calculated inter-rater agreement for each of the four faces separately (see Table 4). Due to missing responses, five participants' data were not included in these calculations.

While Cronbach's α is typically reported as a measure of reliability among raters, especially within the social evaluation literature (e.g., Little & Perrett, 2007; Oosterhof & Todorov, 2008; Zebrowitz, Montepare, & Lee, 1993), the main criticism with this measure is that simply increasing the number of raters produces an increase in the resulting agreement value (Cortina, 1993). As such, we considered five different measures of inter-rater agreement to provide a more complete picture.

Cronbach's α provides a measure of the reliability of the average rater in terms of consistency only (i.e., absolute agreement is ignored). In contrast, the intraclass correlation coefficient, ICC(*A*,*k*), takes into account this absolute agreement by incorporating any systematic differences between raters in terms of the absolute ratings they give. The 'average leave one out' measure quantifies how much we can expect any individual, on average, to agree with the rest of the sample of raters. Kendall's *W* (Kendall, 1948; Kendall & Smith, 1939), a nonparametric statistic, is proportional to the average rank-order correlation among all pairs of raters. Finally, the 'average inter-rater agreement' is simply the average correlation among every possible pair of raters. For more details regarding these measures, see Kramer, Mileva, and Ritchie (2018).

As Table 4 illustrates, taken together, the five measures suggest a medium level of agreement among raters. These values are comparable, for example, with ratings of facial trustworthiness and dominance reported in previous work (Kramer et al., 2018), and can be interpreted as demonstrating a combination of both shared (agreeing with others) and private (idiosyncratic) perceptions. While there is no agreed upon threshold for what is an acceptable level of agreement for facial descriptions, it is clear that an average correlation between pairs of raters in the range of .37 to .48, for instance, is not sufficient if we are to use descriptions to convey facial appearance to others.

It is interesting to note that, in the majority of cases, values of agreement were lower when rating the two female faces in comparison with the two men. While a larger sample of faces would be required before any such pattern can be confirmed, the suggestion is that describing women's faces may be more subjective/private and hence could result in less useful descriptions in terms of eyewitness accounts.

For the yes/no responses in the FRS, we calculated frequencies for each of the four faces separately (see Table 4, lower half). Perhaps surprisingly, a lack of agreement was evident in some of these questions also. For example, raters appeared to disagree on the definition of 'clean-shaven', as well as the presence/absence of dimples for both male and female faces. With these descriptors showing less than perfect agreement, our results highlight a variety of difficulties when a witness is required to describe a face to someone else.

5 General discussion

Our aim in the current work was to investigate the utility of facial descriptions after removing any demands based upon memory. Even without the need to recall a previously seen face, how useful is the description of a face that we provide to others? Our results demonstrated that facial descriptions were far less useful than researchers and professionals might have previously thought.

In Experiment 1A, we found that simply describing a face to a partner for use in selecting that identity from a lineup was highly error-prone, in line with previous research (Wilson, Seale-Carlisle, & Mickes, 2018). When the same image appeared in the lineup as was being described, accuracy in selection was only 73%. However, if the lineup image was different from the one being described (as would almost certainly be the case in real-world scenarios), accuracy dropped to 22%. Finally, when the identity was absent from the lineup, listeners incorrectly chose a face on 46% of trials. These values are concerning, to say the least, if we consider how often witnesses and professionals are required to convey facial appearance to another person during investigations. Facial descriptions provided by witnesses, for example, might produce low levels of correct identifications in combination with high levels of misidentifications. It is worth noting that, importantly, these levels of performance exclude the additional, detrimental effect of memory.

We also found limited evidence to suggest that this difficulty in conveying facial descriptions was the result of difficulties in face matching (Experiment 1B). Previous research has shown that deciding whether two photographs of unfamiliar faces depict the same person or not is prone to error, and that this is true even in the absence of memory demands (Megreya & Burton, 2008). Logically, if one person cannot tell whether two different images show the same face, it is not surprising that a description of one image fails to provide the necessary information to facilitate another person selecting the second image. As such, this process might fail due to both difficulties in describing faces *and* the inability to cope with within-person facial variability (Jenkins, White, Van Montfort, & Burton, 2011).

Our second experiment focussed on agreement in facial descriptions across raters. If people disagree in their descriptions of a particular face then it is hard to see how these descriptions can prove useful, for instance, when searching for that person. While Experiment 1A found that open-ended, freely generated descriptions were unsuccessful in conveying facial information for identification, Experiment 2 utilised predefined rating scales. Therefore, rather than relying on a describer mentioning the size of the nose etc., the rating scales guaranteed that all describers

considered the same facial features. Even so, we found that agreement was far from perfect. Indeed, it was clear that there were substantial individual differences in how faces were rated, and that this was also true for yes/no responses.

It may be the case that constraining descriptions to a specific, predefined set of facial features could limit their utility. The use of more holistic descriptors (e.g., attractive) might benefit descriptions by taking advantage of the fact that we naturally process faces holistically (e.g., Maurer, Le Grand, & Mondloch, 2002; Richler, Mack, Gauthier, & Palmeri, 2009). However, much like the feature descriptions in the present work, perceptions of social traits also show substantial disagreement across observers (Hönekopp, 2006; Kramer et al., 2018). Future research might consider the use of featural versus holistic descriptions in terms of their utility when conveying facial appearance to others.

Allowing witnesses to provide open-ended, freely generated descriptions also appears problematic. This technique is often used in order to create modal descriptions in mock witness paradigms (Doob & Kirshenbaum, 1973). However, as studies utilising this design have demonstrated, collating descriptions across participants and including only those details mentioned by at least 25% of describers typically results in a description that is absent of any facial appearance information (e.g., Humphries et al., 2012). Research in this area, therefore, also appears to support the idea that there is little agreement in facial descriptions across viewers.

Here, we have considered performance levels in absolute terms, perhaps implying that anything lower than 100% accuracy demonstrates a failure in the process and an argument against using facial descriptions. To provide some context, we might compare the current results with those produced through the use of contemporary composite systems. Often, witnesses work with an operator to construct a facial composite (a visual representation of the perpetrator's face) based on their memory of the target. This process typically begins with a verbal description of the face to produce the initial image, which can then be further refined by the witness, resulting in the final composite. Across several systems, the ability for independent observers to match these composites

to the targets in six-image target-present arrays appeared to be limited, ranging from 31% to 60% correct (Frowd et al., 2005; see also Koehn & Fisher, 1997). Given that these researchers provided the same image in the identification array as the one viewed by the witness initially, we might compare this with our 73% correct in Experiment 1A's 'same' condition. The substantially lower performance with composites is likely due to the two-day gap between witnesses viewing the photograph and their subsequent construction of the composite. It is now well established that this delay results in detrimental effects on recall (e.g., Ellis, Shepherd, & Davies, 1980; Meissner, 2002; van Koppen & Lochun, 1997).

Perhaps a more suitable comparison might be with a procedure in which the target image is present (rather than remembered) during composite construction. Using Pro-Fit, researchers found performance levels of 40% to 65%, again using six-image target-present arrays (Bruce, Ness, Hancock, Newman, & Rarity, 2002). This result suggests that facial descriptions may actually compare favourably with performance using composites when memory is not required in either case. However, these results rely on participants comparing a composite with the initial image it was intended to depict, or describing an image so that listeners could select that same image (our 'same' condition). In fact, neither of these scenarios is likely, given that a witness's view of the target will never be identical to the one seen by others (either at the same time or at a later date). As such, performance in our 'different' condition may be more indicative of real-world outcomes.

In the current work, we focussed on descriptions of own-ethnicity faces. Given that previous research has demonstrated poorer performance for other-ethnicity face memory (Meissner & Brigham, 2001) and matching (Megreya, White, & Burton, 2011), we suggest that further investigation may find that descriptions of other-ethnicity faces prove even less useful in lineup identification. Presumably, this would be because describers are less sensitive to the important features and how they might vary when it comes to other-ethnicity faces. However, one study found that White and Black participants showed little difference in the pattern of features used to describe

White and Black faces, although accuracy of descriptions was not assessed (Ellis et al., 1975). Therefore, this would certainly be an interesting route for future study.

One might also consider the use of video identification procedures, where participants are shown constrained videos of the suspect and foils. Video identification parades have become increasingly popular in the UK, for instance, with two different systems in widespread use: VIPER (Video Identification Parade Electronic Recordings) and PROMAT (Profile Matching). Whether facial descriptions that are generated through viewing face videos, or used when selecting from video lineups, would lead to different levels of performance in comparison with photographs remains an open question for research.

Finally, it is interesting to consider whether our inability to usefully describe faces might also generalise to voices. During a criminal investigation, it is likely that the police will ask the victim or witness to provide a description regarding the voice of a suspect (e.g., Skoog Waller & Eriksson, 2016). Such 'earwitness' descriptions are made frequently by those who have encountered perpetrators under poor visual conditions or when an offence is committed over the telephone. It is also important, therefore, to determine the accuracy of voice descriptions given to the police and how useful these may or may not be in aiding identification.

To conclude, our results highlight the difficulties encountered when required to provide a facial description for use by others. Even without memory demands, we found that descriptions were incapable of facilitating lineup identification at levels suitable for real-world purposes. Further, viewers showed only moderate agreement when describing the same face, suggesting fundamental limitations on the utility of facial descriptions. These findings lead us to recommend extreme caution when employing eyewitness descriptions in the pursuit of suspects or in related contexts.

Data availability statement

The data that support the findings of these experiments are available through the Open Science Framework at https://osf.io/swgkq/?view_only=49ab27176d914a29aca060fcf9ccc012.

References

Beresford, J., & Blades, M. (2006). Children's identification of faces from lineups: The effects of lineup presentation and instructions on accuracy. *Journal of Applied Psychology, 91*(5), 1102-1113.

Brown, C., Lloyd-Jones, T. J., & Robinson, M. (2008). Eliciting person descriptions from eyewitnesses: A survey of police perceptions of eyewitness performance and reported use of interview techniques. *European Journal of Cognitive Psychology*, *20*(3), 529-560.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*(4), 339.

Bruce, V., Ness, H., Hancock, P. J. B., Newman, C., & Rarity, J. (2002). Four heads are better than one: Combining face composites yields improvements in face likeness. *Journal of Applied Psychology, 87*(5), 894-902.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

Davies, G. M. (1981). Face recall systems. In G. M. Davies, H. D. Ellis, & J. W. Shepherd (Eds.), *Perceiving and remembering faces*. London: Academic Press.

Davies, G., Tarrant, A., & Flin, R. (1989). Close encounters of the witness kind: Children's memory for a simulated health inspection. *British Journal of Psychology, 80*(4), 415-429.

Deffenbacher, K. A. (1994). Effects of arousal on everyday memory. *Human Performance, 7*(2), 141-161.

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior, 28*(6), 687-706.

Dekle, D. J., Beal, C. R., Elliott, R., & Huneycutt, D. (1996). Children as witnesses: A comparison of lineup versus showup identification methods. *Applied Cognitive Psychology, 10*(1), 1-12.

DiNardo, L., & Rainey, D. (1991). The effects of illumination level and exposure time on facial recognition. *The Psychological Record, 41*(3), 329-334.

Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups - Partial remembering. *Journal of Police Science and Administration, 1*, 287–293.

Ellis, H. D. (1986). Face recall: A psychological perspective. *Human Learning: Journal of Practical Research & Applications, 5*(4), 189-196.

Ellis, H. D., Deregowski, J. B., & Shepherd, J. W. (1975). Descriptions of white and black faces by white and black subjects. *International Journal of Psychology, 10*, 119-123.

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1980). The deterioration of verbal descriptions of faces over different delay intervals. *Journal of Police Science & Administration, 8*, 101–106.

Fallshore, M., & Schooler, J. W. (1995). The verbal vulnerability of perceptual expertise. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*(6), 1608–1623.

Frowd, C. D., Carson, D., Ness, H., McQuiston-Surrett, D., Richardson, J., Baldwin, H., & Hancock, P. (2005). Contemporary composite techniques: The impact of a forensically-relevant target delay. *Legal and Criminological Psychology, 10*(1), 63-81.

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance, 32*(2), 199-209.

Humphries, J. E., Holliday, R. E., & Flowe, H. D. (2012). Faces in motion: Age-related changes in eyewitness identification performance in simultaneous, sequential, and elimination video lineups. *Applied Cognitive Psychology, 26*(1), 149-158.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323.

Kebbell, M. R. (2000). The law concerning the conduct of lineups in England and Wales: How well does it satisfy the recommendations of the American Psychology–Law Society? *Law and Human Behavior*, *24*(3), 309-315.

Kendall, M. (1948). *Rank correlation methods*. London: Charles Griffin & Co.

Kendall, M. G., & Smith, B. B. (1939). The problem of m rankings. *Annals of Mathematical Statistics, 10*, 275–287.

Koehn, C. E., & Fisher, R. P. (1997). Constructing facial composites with the Mac-a-Mug Pro system. *Psychology, Crime and Law, 3*(3), 209-218.

Kramer, R. S. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces. *PLoS ONE, 13*(8), e0202655.

Kuehn, L. L. (1974). Looking down a gun barrel: Person perception and violent crime. *Perceptual and Motor Skills, 39*(3), 1159-1164.

Laughery, K.R., Duval, C., & Wogalter, M.S. (1986). Dynamics of facial recall. In H. D. Ellis, M. A. Jeeves, F. Newcombe, & A. Young (Eds.), *Aspects of face processing* (pp. 373-387). Dordrecht: Martinus Nijhoff.

Lindholm, T., & Christianson, S. -Å. (1998). Gender effects in eyewitness accounts of a violent crime. *Psychology, Crime and Law, 4*, 323-339.

Lindsay, R. C. L., Martin, R., & Webber, L. (1994). Default values in eyewitness descriptions: A problem for the match-to-description lineup foil selection strategy. *Law & Human Behavior, 18*, 527–541.

Little, A. C., & Perrett, D. I. (2007). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology, 98*, 111-126.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6*(6), 255-260.

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14*(4), 364-372.

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *Quarterly Journal of Experimental Psychology, 64*(8), 1473-1483.

Meissner, C. A. (2002). Applied aspects of the instructional bias effect in verbal overshadowing. *Applied Cognitive Psychology, 16*, 911–928.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7*(1), 3-35.

Meissner, C. A., Sporer, S. L., & Schooler, J. W. (2007). Person descriptions as eyewitness evidence. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology (Vol. 2): Memory for people* (pp. 1-34). Mahwah, NJ: Lawrence Erlbaum Associates.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087-11092.

Pickel, K. L. (1999). The influence of context on the "weapon focus" effect. *Law and Human Behavior, 23*(3), 299-311.

Pozzulo, J. D., & Warren, K. L. (2003). Descriptions and identifications of strangers by youth and adult eyewitnesses. *Journal of Applied Psychology, 88*(2), 315-323.

Richler, J. J., Mack, M. L., Gauthier, I., & Palmeri, T. J. (2009). Holistic processing of faces happens at a glance. *Vision Research, 49*(23), 2856-2861.

Shaw, J. S., Garven, S., & Wood, J. M. (1997). Co-witness information can have immediate effects on eyewitness memory reports. *Law & Human Behavior, 21*, 503–523.

Skoog Waller, S., & Eriksson, M. (2016). Vocal age disguise: The role of fundamental frequency and speech rate and its perceived effects. *Frontiers in Psychology, 7*, 1814.

Sporer, S. L. (1992, March). *An archival analysis of person descriptions*. Paper presented at the Biennial Meeting of the American Psychology-Law Society in San Diego, California.

Sporer, S. L. (2007). Person descriptions as retrieval cues: Do they really help? *Psychology, Crime & Law, 13*(6), 591-609.

Steblay, N. M. (1992). A meta-analytic review of the weapon focus effect. *Law and Human Behavior, 16*(4), 413-424.

van Koppen, P., & Lochun, S. (1997). Portraying perpetrators: The validity of offender descriptions by witnesses. *Law and Human Behavior, 21*, 661–685.

Wilson, B. M., Seale-Carlisle, T. M., & Mickes, L. (2018). The effects of verbal descriptions on performance in lineups and showups. *Journal of Experimental Psychology: General, 147*(1), 113-124.

Yarmey, A. D. (1986). Verbal, visual, and voice identification of a rape suspect under different levels of illumination. *Journal of Applied Psychology, 71*(3), 363-370.

Yarmey, A. D. (1993). Adult age and gender differences in eyewitness recall in field settings. *Journal of Applied Social Psychology, 23*(23), 1921-1932.

Yarmey, A. D., Jacob, J., & Porter, A. (2002). Person recall in field settings. *Journal of Applied Social Psychology, 32*(11), 2354-2367.

Yuille, J. C., & Cutshall, J. L. (1986). A case study of eyewitness memory of a crime. *Journal of Applied Psychology, 71*(2), 291-301.

Zebrowitz, L. A., Montepare, J. M., & Lee, H. K. (1993). They don't all look alike: Individual impressions of other racial groups. *Journal of Personality and Social Psychology, 65*(1), 85-101.

Tables

Table 1. Average performance across participants in Experiment 1A, separated by condition.

	Same	Different	Absent
Hits	0.73 [0.67, 0.78]	0.22 [0.19, 0.26]	-
Misidentifications	0.12 [0.09, 0.15]	0.33 [0.28, 0.37]	-
Misses	0.15 [0.12, 0.19]	0.45 [0.40, 0.50]	-
Correct rejections	-	-	0.54 [0.49, 0.59]
False positives	-	-	0.46 [0.41, 0.51]

Note. 95% confidence intervals are given in square brackets.

Table 2. The average frequencies (out of 8 trials) with which each of the features was mentioned in Experiment 1A, separated by condition.

	Same	Different	Absent
Eyes	6.31 (1.31)	6.48 (1.49)	6.43 (1.30)
Nose	3.14 (2.27)	3.15 (2.33)	3.02 (2.28)
Mouth	4.85 (1.87)	5.68 (1.84)	5.52 (2.03)
Hair	6.91 (0.95)	7.66 (0.64)	7.48 (0.79)
Ears	2.09 (1.73)	2.45 (2.00)	2.12 (1.80)
Face shape	2.77 (1.92)	3.00 (2.10)	2.83 (1.95)
Age	1.92 (1.89)	2.14 (1.89)	2.83 (1.87)
Expression	2.72 (2.17)	2.92 (2.46)	2.68 (2.06)

Note. Standard deviations are given in brackets.

Table 3. Average performance across participants in Experiment 1B.

	Mean Proportion
Hits	0.55 [0.51, 0.59]
Misidentifications	0.23 [0.19, 0.27]
Misses	0.22 [0.18, 0.25]
Correct rejections	0.57 [0.52, 0.61]
False positives	0.43 [0.39, 0.48]

Note. 95% confidence intervals are given in square brackets.

Table 4. A summary of the results for Experiment 2.

	ID 1	ID 2	ID 3	ID 4
Sex of face	Male	Male	Female	Female
Number of raters	9	8	10	8
Cronbach's α	0.85	0.87	0.85	0.82
ICC(*A*,*k*)	0.84	0.87	0.85	0.82
Average leave one out	0.60	0.65	0.56	0.56
Kendall's *W*	0.52	0.57	0.46	0.44
Average inter-rater agreement	0.41	0.48	0.37	0.37
Clean-shaven	50%	90%	-	-
No moustache	90%	100%	-	-
No sideburns	90%	80%	-	-
No full beard	90%	100%	-	-
Not cross-eyed	100%	100%	100%	100%
No baggy eyes	50%	70%	100%	100%
No eye rings	70%	80%	80%	100%
No scar	100%	90%	100%	100%
No dimples	70%	90%	70%	40%
No glasses	100%	100%	100%	100%
No ear ring	100%	100%	100%	100%

Note. ICC = intraclass correlation coefficient.

Figure captions

Figure 1. An illustration of the three conditions for an example identity. Image attributions for both photographs: Eva Rinaldi (Own work) [CC BY-SA 2.0].

Figure 2. Proportion correct for the three conditions in Experiment 1A. Error bars represent 95% confidence intervals.